

BLAST Quick Start

blast-help@ncbi.nlm.nih.gov

BLAST Quick Start

Introduction

Algorithm Basics

- Introduction
- Words & extensions

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Basic Local Alignment Search Tool

Compare protein or nucleic acid sequences to protein or nucleic acid databases

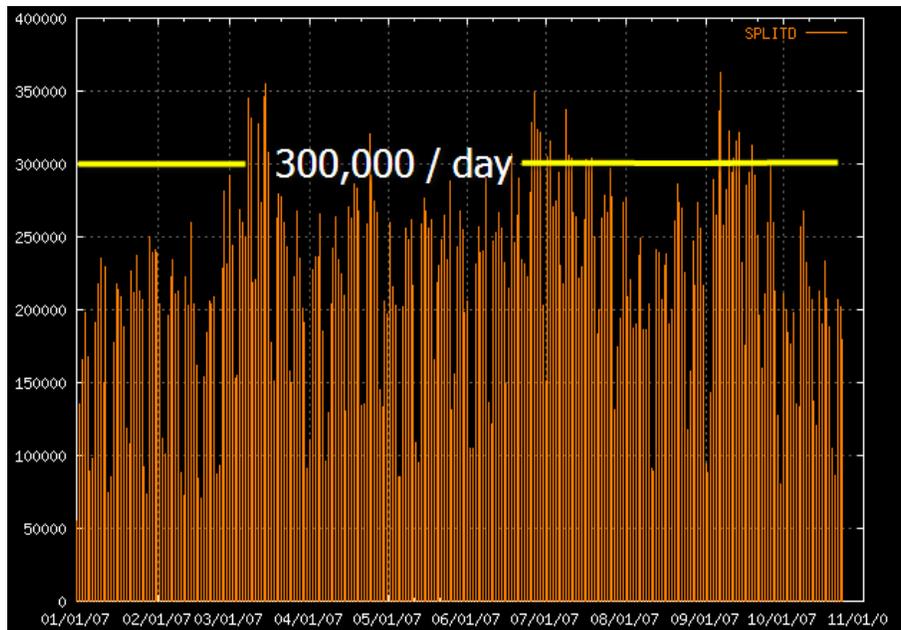
- in NCBI databases
- in local databases (standalone BLAST)
- to a single protein or nucleotide sequence (BLAST 2 Sequences, or pairwise BLAST)

Why do we need similarity searching?

- ◆ To identify and annotate sequences with...
 - incomplete, incorrect, or absent annotations
- ◆ To assemble genomes
- ◆ To explore evolutionary relationships by...
 - finding homologous molecules
 - developing phylogenetic trees

NOTE: Similar sequences may NOT have similar function!

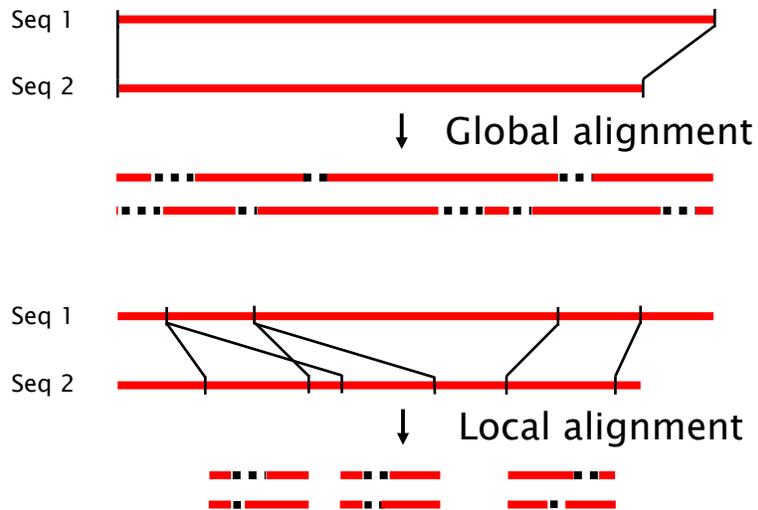
BLAST Web Searches, 2007



Basic Local Alignment Search Tool

- local alignments; isolated regions of similarity
- fast and sensitive
- breaks the query sequence into "words"
- word matches to database sequences are extended in both directions

Global vs Local Alignment



How BLAST Works

1. Make lookup table of “words” for query
2. Scan database for hits
3. Extend alignment both directions

Nucleotide Words

Make a lookup table based on the word size.

11-mer
 ATGCTGCTAGTCGATGACGTAGCTA
 ATGCTGCTAGT
 TGCTGCTAGTC
 GCTGCTAGTCG
 ...

Protein Words

AIEKCYTGCTLAQEADDTA
 AIE
 IEK LEK, IDK, IQK, IER, IDR, etc
 EKC
 Neighborhood words
 KCY
 CYT
 ...

Lookup table, including neighborhood words, is based on word size, score matrix, and threshold.

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

BLAST Programs

What is your goal?

		<u>Word Size</u>
blastn	nucleotide X nucleotide	11
blastp	protein X protein	3
6 frame, translated nucleotide searches		
blastx	nucleotide X protein	3
tblastx	nucleotide X nucleotide	3
tblastn	protein X nucleotide	3

More BLAST Programs

MegaBLAST

- batch nucleotide queries
- very similar sequences

Word Size

28

Discontiguous MegaBLAST

- batch nucleotide queries
- divergent sequences

11 matches
out of 18

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Nucleotide BLAST Databases

- **nr (nt)**
 - Traditional GenBank Divisions
 - NM_ and XM_ RefSeqs
- **refseq_rna**
 - NM_ , XM_ , NR_
- **refseq_genomic**
 - NC_ , NT_ , NG_
- **est**
 - EST Division
- **htgs**
 - HTG division
- **dbsts**
 - STS Division
- **chromosome**
 - NC genomic records
- **gss**
 - GSS division
- **wgs**
 - wgs entries from traditional divisions
- **pdb**
 - Nucleotide sequences from structures
- **env_nt**
 - environmental samples

combined, for human and mouse

New Nucleotide Databases

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastn suite: BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number, gi, or FASTA sequence [Clear](#) Query subrange [From](#) [To](#)

Or, upload file Job Title Entrez query

Choose Search Set

Database

Human genomic + transcript Mouse genomic + transcript Others (nr etc.):

Human genomic plus transcript

Entrez Query Optional

Enter an Entrez query to limit search [?](#)

Human refseq_genomic + refseq_rna = default db

Change Nucleotide Database

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite: BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number

Or, upload file

Job Title

Choose Search Database

Database

Entrez Query Optional

Enter an Entrez query to limit search

Query subrange

From

To

Others (nr etc.):

Genomic plus Transcript

- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases

- Nucleotide collection (nr/nt)
- Reference mRNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- Expressed sequence tags (est)
- Non-human, non-mouse ESTs (est_others)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun reads (wgs)
- Environmental samples (env_nt)

Human genomic plus transcript

Protein BLAST Databases

Protein

- nr
 - traditional GenBank records
- refseq = NP_, XP_
- swissprot
- pdb
- pat
- env_nr

nr ≠ nr

BLAST Databases: Genome-specific

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

[Learn more](#) about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

Human	Oryza sativa	Gallus gallus
Mouse	Bos taurus	Pan troglodytes
Rat	Danio rerio	Microbes
Arabidopsis thaliana	Drosophila melanogaster	Apis mellifera

Basic BLAST

Choose a BLAST program to run.

[nucleotide blast](#) Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast

The screenshot shows the NCBI BLAST website interface. A blue navigation bar at the top contains 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. The main content area is titled 'Basic Local Alignment and Search Tool'. A yellow callout bubble on the left says 'Save your searches' and points to the 'My NCBI' section. A yellow starburst callout on the right says 'What else is new?' and points to the 'BLAST 2.2.13 now available' news item. The interface includes a search bar, a list of species genomes, and various BLAST program options like 'nucleotide blast', 'protein blast', 'blastx', 'tblastn', and 'tblastx'. A 'Tip of the Day' section is visible on the right side.

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Setting up a BLAST Search

The screenshot shows the NCBI BLAST search interface. At the top, there are navigation tabs: Home, Recent Results, Saved Strategies, and Help. Below these is the text "Basic Local Alignment". The main heading is "NCBI/BLAST/blastp suite: BLASTP programs search protein databases using a protein query." Below this, there are several input fields and a dropdown menu. Three yellow callout boxes point to specific elements: "Identifier or sequence" points to the "Enter accession number" field containing "NP_032294"; "Title" points to the "Job Title" field containing "NP_032294:homeo box B5 [Mus musculus]"; and "Select database" points to the "Database" dropdown menu which is open, showing a list of options including "Non-redundant protein sequences (nr)".

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST/ Formatting Results - 41Z6F6C012 [Formatting options]

Job Title: NP_032294:homeo box B5 [Mus musculus]

Putative conserved domains have been detected, click on the image below for detailed results.

homeodonain

WAITING

Request ID: 41Z6F6C012
 Status: Searching
 Submitted at: Thu Mar 22 00:36:54 2007
 Current time: Tue May 8 18:42:35 2007
 Time since submission: 00:00:11

Conserved Domain search run for protein queries

Default Output

NCBI Minicourses

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp/ Formatting Results - 6B1GYW57014 [Reformat these Results] [Edit and Resubmit] [Sign in above to save your search]

Job Title: NP_032294:homeo box B5 [Mus musculus] [Show Cr](#)

BLASTP 2.2.16 (Mar-25-2007)

Reference:
 Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.

Reference:
 Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", *Nucleic Acids Res.* 29:2994-3005.

RID: 6B1GYW57014

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
 5,002,253 sequences: 1,729,200,953 total letters

If you have any problems or questions with the results of this search please refer to the [BLAST FAQs](#)
[Taxonomy reports](#)

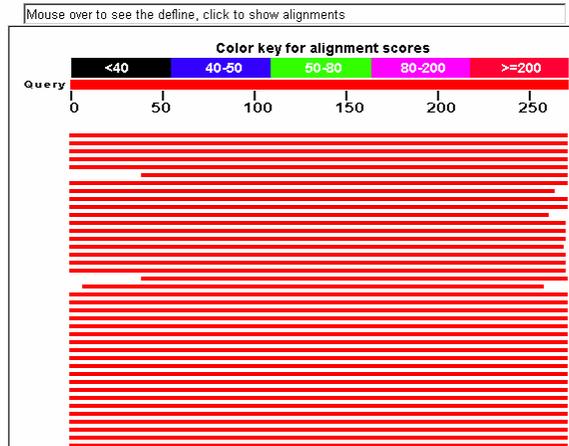
Query= gi|6680251|ref|NP_032294.1| homeo box B5 [Mus musculus]
 Length=269

header

Default Output con't

Graphical Overview

Distribution of 104 Blast Hits on the Query Sequence



Default Output con't

One line descriptions (default view)

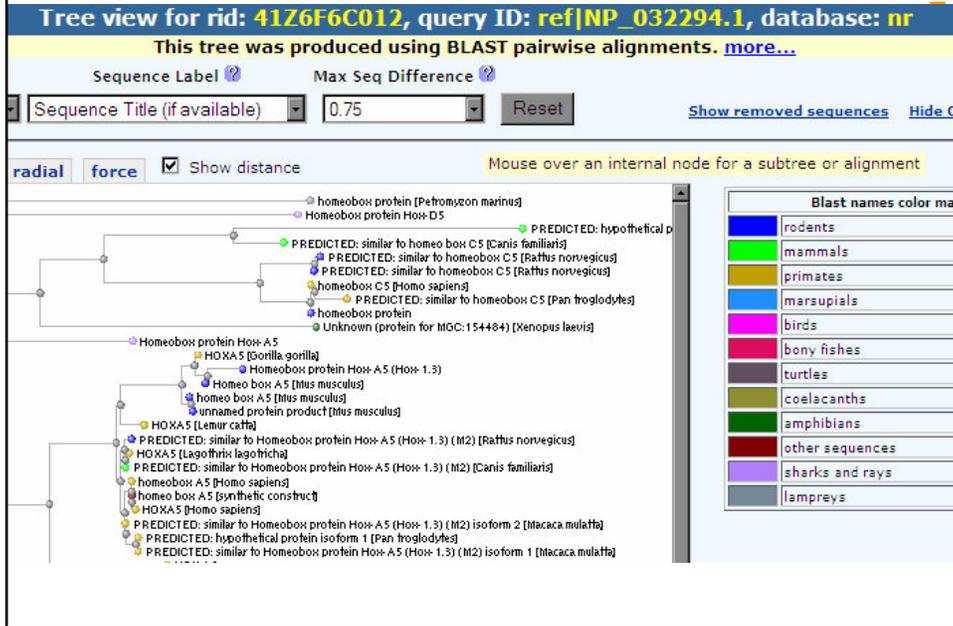


[Distance tree of results](#) NEW

Sequences producing significant alignments:	Score (Bits)	E Value
ref NP_032294.2 homeo box B5 [Mus musculus] >gi 62656894 ref...	553	4e-156
gb AAA37842.1 homeobox protein	551	1e-155
ref XP_548176.2 PREDICTED: similar to Homeobox protein Hox-B...	548	1e-154
ref XP_001502124.1 PREDICTED: hypothetical protein [Equus caballus]	546	4e-154
ref NP_002138.1 homeobox B5 [Homo sapiens] >ref XP_001173004...	546	7e-154
dbj BAB28059.1 unnamed protein product [Mus musculus]	474	3e-132
ref XP_001367145.1 PREDICTED: similar to homeobox protein [Monodelphis domestica]	462	8e-129
ref NP_001020526.1 homeo box B5 [Gallus gallus] >gb AAW48484...	431	3e-119
dbj BAD95556.1 Hoxb-5 [Gallus gallus]	417	4e-115
ref NP_571176.2 homeo box B5a [Danio rerio] >sp P09014 HXB5A...	394	3e-108
emb CAA48320.1 homeodomain protein [Danio rerio] >emb CAA312...	392	1e-107
dbj BAD95561.1 Hoxb-5 [Pelodiscus sinensis]	385	2e-105

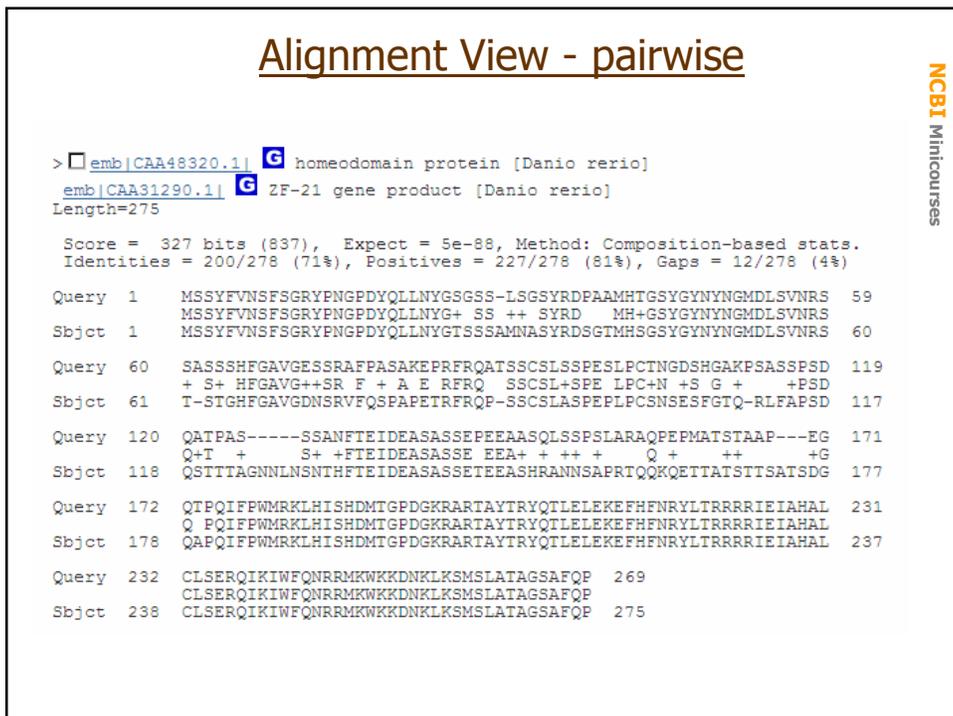
BLAST Output – tree view

NCBI



Alignment View - pairwise

NCBI Minicourses



Scoring Matrices

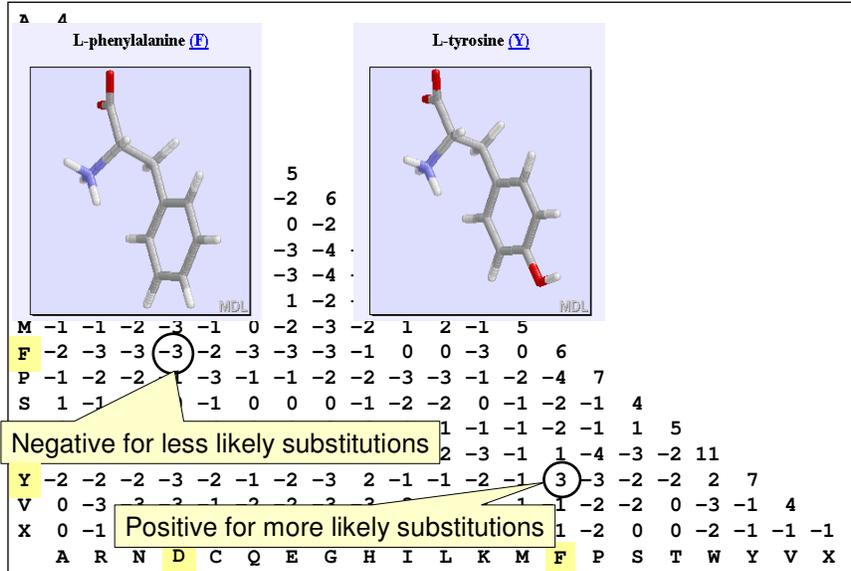
Protein = substitution matrix

- Percent Accepted Mutation (PAM)
- Blocks Substitution Matrix (BLOSUM)

BLOSUM Matrices

- local alignments
- all matrices based on observed alignments; not extrapolated
- BLOSUM 62 calculated from sequences with no more than 62% identity
- Examples: BLOSUM45, BLOSUM62 and BLOSUM80
- BLOSUM62: very good for detecting weak protein similarities
- BLOSUM62 is the default matrix in BLAST

BLOSUM62



Where do Expect Values Come From?

Distance tree of results [NEW](#)

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value
NP_032294.2	homeo box B5 [Mus musculus] >ref[XP_573183.1] PREDICTED: similar to h	553	553	100%	4e-156
AAA37842.1	homeobox protein	551	551	100%	1e-155
XP_548176.2	PREDICTED: similar to Homeobox protein Hox-B5 (Hox-2A) (HHO.C10) (HU-	548	548	100%	1e-154
XP_001502124.1	PREDICTED: hypothetical protein [Equus caballus]	546	546	100%	4e-154
NP_002138.1	homeobox B5 [Homo sapiens] >ref[XP_001173004.1] PREDICTED: homeob	546	546	100%	7e-154
BAB28059.1	unnamed protein product [Mus musculus]	474	474	85%	3e-132
XP_001367145.1	PREDICTED: similar to homeobox protein [Monodelphis domestica]	462	462	100%	8e-129
NP_001020526.1	homeo box B5 [Gallus gallus] >gb[AAW48484.1] homeodomain transcriptio	431	431	100%	3e-119
BAD95556.1	Hoxb-5 [Gallus gallus]	417	417	97%	4e-115
NP_571176.2	homeo box B5a [Danio rerio] >sp[P09014]HXBSA_BRARE Homeobox protei	394	394	100%	3e-108
CAA48320.1	homeodomain protein [Danio rerio] >emb[CAA31290.1] ZF-21 gene produc	392	392	100%	1e-107
BAD95561.1	Hoxb-5 [Pelodiscus sinensis]	385	385	96%	2e-105

Expect Value

E = number of database hits you expect to find by chance, $\geq S$

$$E = Kmne^{-\lambda S} \quad \text{or} \quad E = mn2^{-S'}$$

K = scale for search space

λ = scale for scoring system

$\lambda S'$ = bitscore = $(\lambda S - \ln K) / \ln 2$

m = query length

n = database length

E is dependent on m x n (search space)

More info: [The Statistics of Sequence Similarity Scores](#)

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Initial Options – Search Set

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): mammals (taxid:40674)
Mammalia (taxid:40674)
placental mammals (taxid:9347)
eutherian mammals (taxid:9347)
egg-laying mammals (taxid:9255)
mammalian hepatitis B-type viruses (taxid:10405)
Mammalian virus group (taxid:353212)
Mammalian orthoreovirus (taxid:351073)
Mouse mammary tumor virus (taxid:11757)
Mammillaria (taxid:130139)

Entrez Query (Optional):

Program Selection

Algorithm: BLAST

Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

NCBI Minicourses

Initial Options – Search Set

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query (Optional): Enter an Entrez query to limit search

Program Selection

Algorithm: blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
Choose a BLAST algorithm

BLAST

Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

all[filter] NOT mammalia[organism]

NCBI Minicourses

General Parameters

Max target sequences 100
Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold 10

Word size 3

Scoring Parameters

Matrix BLOSUM62

Gap Costs Existence: 11 Extension: 1

Compositional adjustments Conditional compositional score matrix adjustment

Filters and Masking

Filter Low complexity regions

Mask Mask for lookup table only
 Mask lower case letters

Advanced Options

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Reformat

NCBI Minicourses

► NCBI BLAST/ blastp/ Formatting Results - N5W6JP2F013 [Reformat these Results] [Edit and Resubmit] [Sign in]

Job Title: NP_032294:homeo box B5 [Mus musculus]

BLASTP 2.2.17 (Aug-26-2007)

Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.

Reference:

Stephen F. Altschul, John C. Wootton, E. Michael Gertz, Richa Agarwal, Aleksandr Morgulis, Alejandro A. Schäffer, and Yi-Kuo Yu "Protein database searches using compositionally adjusted substitution matrices", *FEBS J.* 272:5101-5109.

RID: N5W6JP2F013

Database: All non-redundant GenBank CDS

translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
5,678,482 sequences; 1,961,803,296 total letters

Reformat

NCBI

Query ref|np_032294| (269 letters)

Database nr

Job title NP_032294:homeo box B5 [Mus musculus]

Request ID N5W6JP2F013

Format

View report Show results in a new window

Reset form to defaults

Advanced View

Alignment View

Pairwise

Graphical Overview

Linkout

Sequence Retrieval

NCBI-gi

Masking Character

Descriptions

Organism

Type code

Enter organism name

Entrez query

Expect Min.

Expect Max.

PSI-BLAST

with inclusion threshold:

Lower Case

X for protein, n for nucleotide

Lower Case

BLAST Output: Low Complexity Filter

NCBI

```
> gi|466462|gb|AAA17374.1 human homolog of E. coli mutL gene product, Swiss-Prot
   Number P23367
   Length=756

Score = 219.935 bits (559), Expect = 8.71148e-57
Identities = 120/131 (91%), Positives = 125/131 (95%), Gaps = 2/131 (1%)

Query 1 IETVYAAALPKNTHFFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILE-VQQHIESKLL 59
      IETVYAAALPKNTHFFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILE VQQHIESKLL
Sbjct 276 IETVYAAALPKNTHFFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 335

Query 60 GSNSSRMYYFTQTLLEGLAGPSGEMVK-sttsltssstsgsDKVYAHQMVRTDSREQKLLDA 118
      GSNSSRMYYFTQTLLEGLAGPSGEMVK +T+ +SS+ SDKVYAHQMVRTDSREQKLLDA
Sbjct 336 GSNSSRMYYFTQTLLEGLAGPSGEMVKSTTSLTSSSTSGSDKVYAHQMVRTDSREQKLLDA 395

Query 119 FLQPLSKPLSS 129 low complexity sequence filtered
      FLQPLSKPLSS
Sbjct 396 FLQPLSKPLSS 406
```

Alignment View Options

NCBI Minicourses

Alignment View

- Pairwise
- Pairwise
- Pairwise with dots for identities
- Query-anchored with dots for identities
- Query-anchored with letters for identities
- Flat query-anchored with dots for identities
- Flat query-anchored with letters for identities
- Hit Table

Alignment View

- Pairwise
- Pairwise
- Pairwise with dots for identities
- Query-anchored with dots for identities
- Query-anchored with letters for identities
- Flat query-anchored with dots for identities
- Flat query-anchored with letters for identities
- Hit Table

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Examples

- Protein searches more sensitive
than nucleotide searches
 - redundancy of the genetic code
- megablast best when searching within
same organism

BLAST is a shortcut . . .

An alignment BLAST cannot make:

```

1 GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG
  ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
1 GAGTGATACGATGAGCCCAGTGTAGCAGTGAAGATCTGGACCACGGTGTACTCGTTGTCC

61 GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAAGAACTCCTGGGGAGCCAGTT
  ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
61 GCTATGGTGTTAAGGGTGGGAAGAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT

121 GGGGTGAACAAGGTTATTTTCAGGCTTGCTCGTGGTAAAAAC
   ||||| ||||| ||| ||| ||| ||| ||| ||| |||
121 GGGGAGACCAAGGCTACATCCTTATGTCCCCTGACAACAAC
  
```

Reason:
no contiguous exact match of 7 bp.

Nucleotide vs. Protein BLAST

Comparing ADSS from *H. sapiens* and *A. thaliana*

	a	a	c	g	g	g	g	a	c	g	g	g	t	g	t	g	t	c	t	c	g	g	t	g	c	a	t	g	g	g	g	g	a	c	a	g	a	g	g	c	
Human:	N	R	V	T	V	V	L	G	A	Q	W	G	D	E	G																										
A.th.:	S	Q	V	S	G	V	L	G	C	Q	W	G	D	E	G																										
	a	g	t	a	a	g	t	a	t	c	t	g	t	g	t	a	t	c	t	c	g	g	t	g	c	a	t	g	g	g	g	a	c	a	g	a	g	g	t		

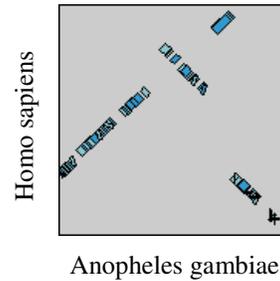
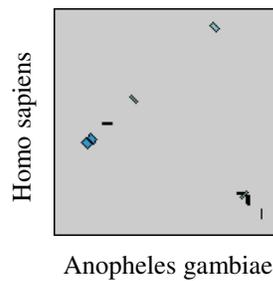
- BLASTp finds three matching words
- BLASTn finds no match, because there are no 7 bp words

Protein searches are generally more sensitive than nucleotide searches.

BLASTN vs TBLASTX

Anopheles gambiae mitochondrion, complete genome.
ACCESSION NC_002084 GI:5834911

Homo sapiens mitochondrion, complete genome.
ACCESSION NC_001807 GI:17981852



Hands-On

www.ncbi.nlm.nih.gov/Class/minicourses/quickblast.html